Supplementary Note

Mouse Segmental Duplication and Copy-Number Variation

Xinwei She[1], Ze Cheng[1], Sebastian Zöllner[2], Deanna Church[3], Evan E. Eichler[1,4]

I. Segmental Duplication Analysis.

We analyzed the mouse genome assembly (Build36) using two different genome-wide approaches designed to detect genomic duplicates >90% sequence identity.

Whole Genome Analysis Comparison (WGAC). A BLAST-based whole genome assembly comparison (WGAC) method[1] was used to identify all pairwise alignments >1 kb and >90% identity within the February 2006 mouse (*Mus musculus)* "essentially complete" genome assembly (NCBI Build36). The procedure eliminates all common interspersed repeat families from the initial "seeding" alignments but reintroduces common repeats as part of the optimal global alignment stage in an effort to define the boundaries of segmental duplications more accurately. The analysis identified 138,768 pairwise alignments (>1 kbp in length and >=90% identity) corresponding to 141.4 Mb of the C57BL/6J genome. 86.6 Mb of these belonged to duplications >10 kbp and >94% in length. The latter corresponded to a total of 14,342 pairwise alignments (13,705 mapping within a chromosome and 637 mapping between non-homologous chromosomes).

Whole Genome Shotgun Sequence Detection (WSSD). As larger, high-identity duplications (>94%) are frequently collapsed within working draft sequence assemblies (She et al., 2004), we compared these assembly-based results to whole genome shotgun sequence detection (WSSD) database of mouse segmental duplications. WSSD identifies duplicated regions >10 kb in length based on a significant excess of WGS read depth-of-coverage (WSSD_DOC)[2] or an excess of sequence divergence (WSSD_DRR=divergent read ratio)[3]. Thresholds are established by analysis of 1.9 Mb of unique mouse sequence corresponding to 12 finished BAC clones as described previously[3]. The analysis is performed by aligning all mouse whole-genome shotgun sequence reads (n= 40,782,208) to 400 kb segments of the mouse Build36 assembly. We exclude all common repeats with less than 5% divergence from their consensus as well as LTR and LINE elements with <15% and 10% divergence from their consensus. 24,741,545 WGS reads were remapped and aligned to the assembly based on the following criteria: a minimum of 400 bp of aligned read length; >94% sequence identity; >300 bp non-repeat-masked bp and at least 200 bp of high quality sequence (Phred Q>30). Combining duplication intervals predicted by WSSD_DOC and WSSD_DRR we estimated 131.2 Mb of segmental duplication of which 83.2 Mb was shared between WGAC and WSSD. We conservatively estimated the total duplication content based on the union of shared duplications and all WGAC duplications <94% sequence identity.

Supplementary Table 1: Mouse Segmental Duplication Analysis

Table 1: Mouse Segmental Duplication Analysis (Mar. 2006).

chrom	Non gap size	Total WGAC	WGAC >10K, >94%	WSSD_DOC +DRR	Shared	Total	%
Total	2,550,156,572	141,408,316	86,634,191	131,208,151	83,135,830	125,975,847	4.94%

* Excludes sequence assigned to the random bin as well as Y chromosome which was not analyzed as part of this assembly

** Total duplication is the union of WGAC (<94%) and shared based of WSSD and WGAC

Although the genome assembly has improved significantly, there are several indications that the assembly of these regions is not yet complete. First, we have identified 3 Mbp of whole-genome assembly comparison duplications that can not be confirmed by WSSD, and conversely 38 Mbp that show evidence of duplication by WSSD but are not confirmed by WGAC. We find that the gap regions (similar to the human genome) are particularly enriched in segmental duplications with 23% of the gaps are flanked by duplications (as detected by both measures). Interestingly, 10 gap regions are flanked by duplications using the WSSD method alone, suggesting that these regions may represent uncharacterized duplications within the assembly. Second, if we compare an optical restriction map of the mouse genome assembly, we find that these regions are enriched in mapping discrepancies (Church et al., unpublished). Once again these inconsistencies are restricted to local regions embedded within duplications and most likely reflect incomplete assembly or variation within large tandem duplications.

II. Copy-Number Variant Detection.

Combining information across individuals, we identified regions of copy-number variation from the observed hybridization signal using a novel hidden Markov type method. We model copy-number at each probe as one of three hidden states relative to C57BL/6J: state 0 = no difference, state 1 = copy gain and state 2 = copy loss. Dependent on each probe's state, the hybridization signal is generated from a mixture of normal distributions with variance 1 and average M. The states of n consecutive probes $s_1,\ldots,s_n$ can be considered a Markov chain. Then the transition probability $tr(s_i,s_{i+1})$ between consecutive probes s_i, s_{i+1} depends on their location relative to CNVs in the sample. If no CNV boundaries overlap with the region between s_i and s_{i+1}, then $tr(s_i=0,s_{i+1}=0)=1$. However, if the proximal boundary of, for example, a copy gain with population frequency f is located between consecutive probes i and $i+1$, the transition probabilities are

$$tr(s_i = 0, s_{i+1}) = \begin{cases} f & if \;\; s_{i+1} = 1 \\ 1-f & if \;\; s_{i+1} = 0 \\ 0 & if \;\; s_{i+1} = 2 \end{cases}.$$

Note that other than in standard hidden Markov models, the transition matrix varies between every pair of markers; here transition probabilities $tr(s_i=0,\ s_{i+1}=1/\ s_{i+1}=2)>0$ indicate boundaries of CNVs.

Each individual in the sample is one realization of this Markov process. Thus we can use all individuals jointly to estimate each transition matrix. The transition matrix at every position is sufficient to identify the location and the frequency of all CNVs segregating in

the population. To estimate all $tr(s_i,s_{i+1})$ we use the Baum-Welch algorithm. As the probes on our chip are tightly spaced, we expect all CNVs to span multiple probes. Therefore, we extend the underlying Markov model by requiring that each state is unchanged for at least 24 consecutive probes, setting a minimum length for each CNV (~12 kb). We do not estimate the mean signal intensity of the gain/loss state, rather we set M_1 and M_2 such that the number of false positives in a simulated dataset of 100,000 probes is 0 ($M_1=2$, $M_2=-2$). After performing an iteration of the forward-backwards algorithm for each individual, we obtain estimates of the transition probabilities for every adjacent pair of probes for each individual. Averaging these probabilities over all individuals, we estimate the joint transition matrix from all individuals and thus combine the evidence for copy-number variation across individuals. As the second step, we use the Baum-Welch algorithm while keeping the transition probabilities constant, thus estimating M and the state probability of each probe $Pr(s_i^{(k)})$in each individual k. To identify regions harboring CNVs from the estimated parameters, we consider pairs of probes with $tr(s_i=0,s_{i+1}=1/s_{i+1}=2)>0.05$ as a proximal boundary of a copy gain/copy loss. To locate the terminal boundary, we sum the transition probabilities back to the baseline state

$$\sum_{j=i+1}^{m} tr(s_j = 1 / s_j = 2, s_{j+1} = 0)$$

and call as the last probe of the CNV the probe s_m where this cumulative transition probability exceeds 0.95. For each individual k, the probabilities of carrying a CNV are calculated by averaging the state probabilities $Pr(s_{i+1}^{(k)}=1/ s_{i+1}^{(k)}=2),\ldots, Pr(s_m^{(k)}=1/ s_m^{(k)}=2)$ over all probes covered by the CNV.

This algorithm is implemented in the program CopyMap, which is available for a free download at http://www.sph.umich.edu/csg/szoellner/software/.

As a second approach for copy-number variation, we developed a simple heuristic-based approach to identify sites of copy-number variation based strictly on the $\log_2$ relative hybridization signal intensity differences between C57BL/6J and the test genome. Our second approach simply identified all windows where 40 adjacent probes showed an average $\log_2$ signal intensity >1 SD beyond the mean when compared to unique control regions within the mouse genome. Overlapping windows were concatenated to generate a CNV interval. We compared the HMM CNV calls versus the heuristic CNV intervals and found good agreement, although the heuristic based approach consistently predicted an additional 3%-5% of the examined basepairs as copy-number variant.

Supplementary Table 2: Heuristic vs. HMM: Genotype calls.

Mouse CNV detected by heuristic algorithm

		Losses				Gains				All polymorphic sites (Gains or loss)				Regions with both gains and losses	
Regions	Length of Reions (bp)	Average (bp/strain)	Fraction	redundant space in all strains (bp)	Fraction	Average (bp/strain)	Fraction	Non redundant space in all strains (bp)	Fraction	Average (bp/strain)	Fraction	Non redundant space in all strains (bp)	Fraction	Non redundant space in all strains (bp)	Fraction
SD	97,861,941	15,664,770	16.0%	52,337,597	53.5%	9,787,565	10.0%	43,843,124	44.8%	25,452,336	26.0%	74,885,601	76.5%	31,042,477	31.7%
SD + 10k flanking	120,797,398	16,756,260	13.9%	57,165,128	47.3%	10,632,425	8.8%	48,363,193	40.0%	27,388,685	22.7%	82,945,089	68.7%	34,581,896	28.6%
10kb falnking	22,935,457	1,091,489	4.8%	4,827,529	21.0%	844,859	3.7%	4,520,066	19.7%	1,936,348	8.4%	8,059,483	35.1%	3,539,417	15.4%
Cai_cnp	49,705,715	4,345,680	8.7%	12,414,813	25.0%	2,674,975	5.4%	8,954,061	18.0%	7,020,655	14.1%	16,432,539	33.1%	7,478,478	15.0%
Cai_cnp (non SD)	37,942,601	940,842	2.5%	2,392,600	6.3%	708,397	1.9%	2,802,698	7.4%	1,649,239	4.3%	4,821,009	12.7%	2,018,311	5.3%
All probe regions	159,423,583	17,567,069	11.0%	59,378,752	37.2%	11,242,699	7.1%	51,055,125	32.0%	28,809,768	18.1%	87,569,461	54.9%	36,514,336	22.9%

Mouse CNV detected by Hiden Markov Model algorithm

		Losses				Gains				All polymorphic sites (Gains or loss)				Regions with both gains	
Regions	Length of Reions (bp)	Average (bp/strain)	Fraction	redundant space in all strains (bp)	Fraction	Average (bp/strain)	Fraction	Non redundant space in all strains (bp)	Fraction	Average (bp/strain)	Fraction	Non redundant space in all strains (bp)	Fraction	Non redundant space in all strains (bp)	Fraction
SD	97,861,941	12,842,070	13.1%	38,956,049	39.8%	7,243,535	7.4%	30,054,658	30.7%	20,085,606	20.5%	60,290,766	61.6%	8,719,941	8.9%
SD + 10k flanking	120,797,398	13,881,269	11.5%	42,948,451	35.6%	8,016,974	6.6%	33,432,606	27.7%	21,898,243	18.1%	66,997,649	55.5%	9,383,408	7.8%
10kb falnking	22,935,457	1,038,993	4.5%	3,992,402	17.4%	773,301	3.4%	3,377,948	14.7%	1,812,295	7.9%	6,706,883	29.2%	663,467	2.9%
Cai_cnp	49,705,715	3,340,271	6.7%	8,659,088	17.4%	2,092,868	4.2%	6,095,035	12.3%	5,433,140	10.9%	13,077,040	26.3%	1,677,083	3.4%
Cai_cnp (non SD)	37,942,601	643,436	1.7%	1,941,020	5.1%	495,423	1.3%	1,701,915	4.5%	1,138,859	3.0%	3,486,798	9.2%	156,137	0.4%
All probe regions	159,423,583	14,466,209	9.1%	44,898,405	28.2%	8,466,401	5.3%	35,061,663	22.0%	22,932,610	14.4%	70,462,967	44.2%	9,497,101	6.0%

Next, we compared both of these methods to 42 “high confidence” copy-number variants in intervals that had been predicted previously by Graubert and colleagues in 5 inbred strains that overlapped with our dataset. The comparison showed that both approaches were comparable and performed well correctly identifying 95% of these high-confidence sites; 41/42 sites were confirmed using the adjacency average approach while the HMM approach confirmed similarly 41/42 sites (the two discordant sites differed using the two different methods). As a control, we performed an analysis using two different individuals from the C57BL/6J to provide an estimate of false positives using these approaches. Using the heuristic approach, we identified 36/913 regions as CNV based on this self-comparison. In contrast, we identified a total of 4/2424 intervals of potential copy-number variation using the HMM. Two of these corresponded to the IgH region on mouse chromosome 12—a region of known somatic instability. Correcting for these exceptions, we find two sites of variation between these isogenic individuals suggesting that the HMM is sufficiently robust. For simplicity, we report all subsequent analyses using the HMM approach. We recognize, however, that conservative nature of the HMM tends to overfragment CNV regions leading typically to twice the number of regions when compared to the heuristic-based approach.

Supplementary Table 3: Heuristic vs. HMM: Genotype calls.

Comparison of CNVs identified by the high density probe array and previously reported CNV (Graubert et all, 2007)

Comparison with Graubert et all, 2007

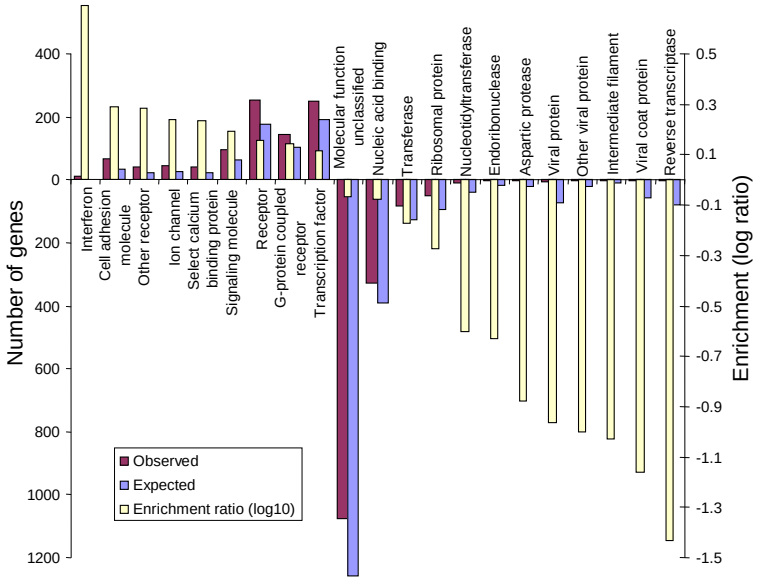
	DBA/2J		A/J		CAST/EiJ		BALB/cByJ		NOD/LtJ	
CNV calling algorithm	Heuristic	HMM	Heuristic	HMM	Heuristic	HMM	Heuristic	HMM	Heuristic	HMM
Graubert CNV covered by probes	11		3		6		13		14	
Graubert CNV not covered by probes	7		3		2		5		13	
Concordant	11	11	3	3	5	6	13	13	14	13
Discordant	0	0	0	0	0	1	0	0	2	1
Intersection with Graubert gains	7	23	8	21	12	36	11	32	5	12
Novel gains	91	221	119	269	166	438	173	406	147	346
Intersection with Graubert losses	15	37	19	61	12	46	17	66	19	64
Novel losses	241	638	227	534	226	792	208	543	220	566

III. Gene Ontology Analyses.

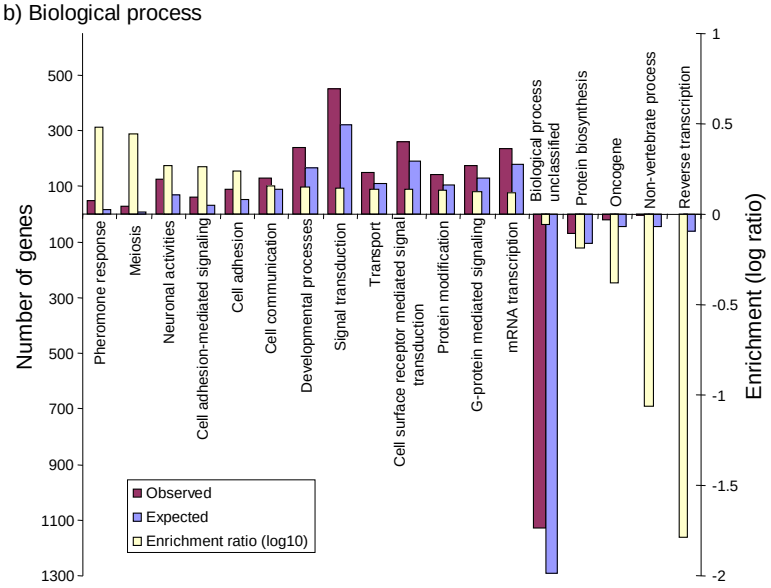
We searched for overrepresentation of gene ontology classifications among RefSeq genes assigned to mouse segmental duplications. RefSeq genes with at least one exon fully contained within segmental duplications were subjected to a gene ontology analysis (http://www.pantherdb.org) based on molecular function and biological process (Fig. 1). We found an enrichment of molecular functions, such as cell adhesion, ion channel, calcium binding protein, signal molecules, receptors and transcription factors, while genes involved in nucleic acid binding, nucleotidyl transferase, reverse transcriptase and other virus packaging are rare. For biological processes, duplicated genes are enriched in meiosis, signal transduction, cell adhesion and transcription (Supplementary Fig. 1). We repeated the analysis for genes fully contained within copy-number variant regions with similar (albeit less significant results) (Table 1). Similar gene enrichments have been observed for both copy-number variant and duplicated regions of other genomes.

Supplementary Figure 1. Overrepresented Gene Classes within Segmental Duplications by a) molecular function and b) biological process. Only the gene categories in which observed gene numbers are significantly different from expected are depicted ($p<0.05$, based on hypergeometric distribution with Bonferroni correction). Gene categories are displayed below the X-axis to fit the negative log ratios, when the number of observed genes is smaller than the expected number of genes.

a) Molecular function



b) Biological process



We repeated this analysis for genes completely and/or partially covered by CNV regions classifying both by biological process and molecular function (http://www.pantherdb.org). The enrichment in each gene ontology classification is based on the background reference of total NCBI Entrez genes. P value of enrichment is calculated by hypergeometric distribution with Bonferonni correction.

Supplementary Table 4: Gene Ontology Analysis for CNV Genes within Segmental Duplications

		Classification	Number of genes	Enrichment	P value
Genes partially or completely covered by CNV regions	Biological Process	MHCI-mediated immunity	16	23.4	1.03E-15
		Immunity and defense	52	3.0	6.61E-10
		Natural killer cell mediated immunity	12	13.9	1.31E-08
		T-cell mediated immunity	17	5.9	1.61E-06
		Gametogenesis	15	5.5	2.67E-05
		Interferon-mediated immunity	10	8.1	1.07E-04
		Spermatogenesis and motility	9	6.1	4.52E-03
	Molecular Function	Defense/immunity protein	38	6.6	1.02E-17
		Major histocompatibility complex antigen	11	14.3	6.91E-08
		Other defense and immunity protein	11	9.9	3.90E-06
		Protease inhibitor	14	6.6	8.65E-06
		KRAB box transcription factor	21	3.7	8.93E-05
		Chemokine	7	14.6	1.12E-04
		Zinc finger transcription factor	25	2.6	4.05E-03
Genes completely covered by CNV regions	Biological process	MHCI-mediated immunity	13	34.7	1.24E-14
		T-cell mediated immunity	14	8.8	2.11E-07
		Immunity and defense	29	3.0	3.04E-05
		Gametogenesis	9	6.1	4.88E-03
		Other oncogenesis	5	12.2	1.36E-02
		Biological process unclassified	57	1.6	1.83E-02
	Molecular Function	Defense/immunity protein	20	6.3	1.88E-08
		Major histocompatibility complex antigen	8	18.9	2.57E-06
		Chemokine	5	18.9	1.63E-03
		Amylase	3	54.0	4.56E-03
		Other defense and immunity protein	6	9.8	8.89E-03
		Molecular function unclassified	53	1.6	3.04E-02

Reference

1. Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J. & Eichler, E.E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res* **11**, 1005-17. (2001).
2. Bailey, J.A. et al. Recent segmental duplications in the human genome. *Science* **297**, 1003-7 (2002).
3. Bailey, J.A., Church, D.M., Ventura, M., Rocchi, M. & Eichler, E.E. Analysis of segmental duplications and genome assembly in the mouse. *Genome Res* **14**, 789-801 (2004).